2013 International Conference on Technology, Informatics, Management, Engineering & Environment (TIME-E 2013)
Bandung, Indonesia, June 23-26, 2013

Content Based Image Retrieval And Classification Using Speeded-Up Robust Features (SURF) and Grouped Bag-of-Visual-Words (GBoVW)

Alexandra Alfanindya, Noramiza Hashim, Chikannan Eswaran
Faculty of Computing and Informatics
Multimedia University
Cyberjaya, Malaysia

Abstract—**This paper presents a work in progress for a proposed method for Content Based Image Retrieval (CBIR) and Classification. The proposed method makes use of the interest points detector and descriptor called Speeded-Up Robust Features (SURF) combined with Bag-of-Visual-Words (BoVW). The combination yields a good retrieval and classification result when compared to other methods. Moreover, a new dictionary building method in which each group has its own dictionary is also proposed. Our method is tested on the highly diverse COREL1000 database and has shown a more discriminative classification and retrieval result.**

Keywords—Speeded-Up Robust Features (SURF), Bag-of-Words (BOW), Bag-of-Visual-Words (BoVW), COREL1000, Content Based Image Retrieval (CBIR)

I. INTRODUCTION

Digital imaging is on the rise in the last few decades. With the internet and World Wide Web (WWW) now widely accessible anywhere and anytime, people are embracing the possibility of accessing images stored thousands of miles away and use it for their own purposes. But retrieving a desired image within a large scale collection with thousands of images is a stressful task. Most image retrieval systems rely heavily on the text based descriptions or annotation [1]. But the text based image retrieval has a heavy limitation in which it relies heavily on manually annotating images one by one. It also depends on the annotator interpretation of the image which can vary from one person to another. Problems with the traditional method of image annotation have led to the rise of interest in techniques for retrieving images based on the content.

Early CBIR system made use of low level visual features such as color and texture. Some early works include the work by M.J. Swain and D. H. Ballard [2] in which they proposed the concept of color histogram as well as introduced the concept of histogram intersection distance metric to measure the distance between the histogram of images. Another early work is the work by S.K Chang and S.H Liu [3] in which abstraction operations are formulated to perform clustering and classification of picture object. Low level visual features are sensitive to factors such as rotation and illumination. Eventhough there are a lot of works trying to fix that, there still exist a 'semantic gap' [4] between low level visual features and the richness of human semantics [5] because of the difference between computer machine and human brains.

Other works that have been done in CBIR have been focusing on narrowing the 'semantic gap' between human and computers. Such works made use of many methods: using object ontology to define high level concept, using machine learning methods to associate low level features with query concepts, using relevance feedback to learn user's intention, generating semantic template to support high level image retrieval, and fusing the evidences from HTML text and the visual content of images for WWW image retrieval [6].

In this research, we utilize a sophisticated way of image feature extraction and indexing using SURF and BoVW. SURF algorithm [7], or Speeded-Up Robust Features, is a robust image local features detector which detects interest points and produces their descriptors. The interest points are not only distinctive, but also robust to noise, detection errors, as well as geometric and photometric changes. Interest points are key points that have well-defined locations in image scale space. They roughly represent the object of the image. Meanwhile, BoVW is the computer vision application of Bag-of-Words (BoW) model for text retrieval that assumes text documents as an unordered collection of words. The BoW model will be further explained in Chapter II.

Our proposed method which we call Grouped BoVW (GBoVW) is different with the normal BoVW. The normal BoVW only has 1 global dictionary and our GBoVW has a dictionary for each group or class in our test database, which make our method more discriminative and results in higher accuracy.

This paper is organized as follows: Chapter II describes the algorithm used. Chapter III describes how the system is built and also the experiment setup. Chapter IV discusses the result of the experiment. Chapter V draws conclusion from all of the experiment.

II. ALGORITHMS

A. Speeded-Up Robust Features (SURF)

Herbert Bay et. al. [7] first introduced the SURF algorithm as a novel scale- and rotation-invariant interest point detector

978-1-4673-5732-6/13/$31.00 ©2013 IEEE

and descriptor. SURF produces a set of interest points for each image and a set of 64-dimensional descriptors for each interest points.

To detect interest points, SURF algorithm is based on the Hessian Matrix, but uses a very basic accurate approximation of Hessian determinant using the Difference-of-Gaussian (DoG). DoG is a very basic Laplacian-based detector. The descriptor uses a distribution of Haar-wavelet responses around the interest point's neighborhood.

SURF algorithm is very similar to SIFT algorithm [8], introduced by David G. Lowe, in term that they are both an interest points detector and descriptors as image features. In SIFT, these features are identified by using a staged filtering approach. The first stage identifies key locations in scale space by looking for locations that are maxima or minima of a Difference-of-Gaussian (DoG) function. Each point is used to generate a feature vector that describes the local image region sampled relative to its scale-space coordinate frame.

The major difference between SIFT and SURF is that, in the implementation of scale-space, SIFT typically implemented image pyramid where the input image is iteratively convolved with Gaussian kernel and repeatedly sub-sampled (reduced in size) [9]; while SURF created scale-space by applying kernels of increasing size to the original image. Another difference is that SURF descriptor has 128 dimensions while SURF descriptor only has 64 dimensions. Some comparison papers such as [10], [11], and [12] have stated that SURF outperforms SIFT in terms of result and computational time, thus we chose SURF instead of SIFT as our feature extractor.

SURF has 4 major steps as explained in [9] and [13]:

1. Integral Image
    - Creates the integral image representation of supplied input image.
    - Calculates pixel sums over upright rectangular areas.

2. Fast Hessian
    - Builds the determinant of Hessian response map.
    - Performs a non-maximal suppression to localize interest points in a scale-space resulting in vector of localized interest point.
    - Uses the determinant of Hessian Matrix

$$H\big(f(x,y)\big) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x^\partial y} \\ \frac{\partial^2 f}{\partial x^\partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} \qquad (1)$$

$$\det(H) = \frac{\partial^2 f}{\partial x^2}\frac{\partial^2 f}{\partial y^2} - \left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 \qquad (2)$$

    - Interpolates detected points to sub-pixel accuracy.

3. SURF Descriptor
    - Calculates dominant orientation of the interest points.
    - Constructs a 4x4 window around the interest point.
    - Calculates Haar Wavelet responses from each sub-region at 5x5 regularly-spaced sample points.
    - Extracts 64-dimensional descriptor vector based on sums of wavelet responses.

4. Salient Features
    - Stores data associated with each individual interest point.

Figure 1 shows an example of SURF Interest points in image number 414 (Dinosaur) from COREL1000 database:

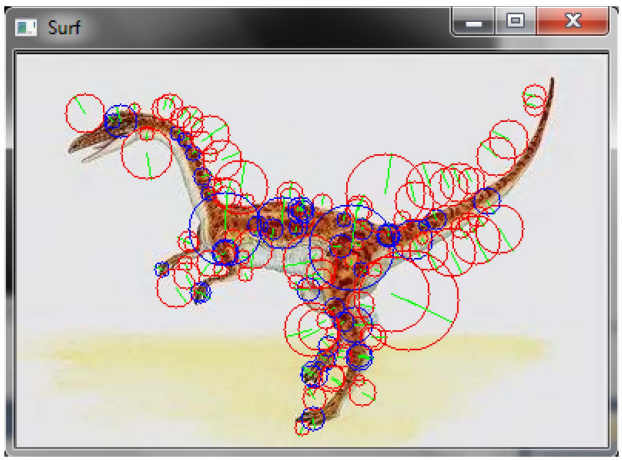


Figure 1. Example of SURF Interest Points

B. Bag-of-Visual-Words (BoVW)

Bag-of-Words (BoW) was originally devised as a text document retrieval algorithm. It describes a document based on the words it contained and the frequency of the word appearance. So the BoW considers "John loves Jane" the same as "loves Jane John" because both contains the same 3 words and the same frequencies of occurrences.

In [8], the BoVW approach was first tried out by clustering SIFT features introduced for object recognition. Ke Gao et. al. [14] proposed a filtration for SIFT interest points using attention model and then used BoVW model to efficiently index the filtered interest points. Pedro Quelhas et. al [15] use BoVW (in the paper they call it Bag-of-Visterms) to index and classify scenery images with DoG interest points. Tom Botterill, Steven Mills, and Richard Green [16] used the combination of SURF and Normal BoVW (with global

dictionary) for robot localization through scene recognition. Anne Bosch [17] et al concluded in their review paper that BoW method achieved the best classification result for scenery classification.

Generally, the BoW consists of 3 main steps:

1. Automatically extract the interest points and descriptor from the images.
2. Quantize the keypoints and descriptors to form the visual dictionary.
3. Find the occurences of each visual words in the image in order to build the BoW histogram.

III. EXPERIMENTAL SETUP

The prototype was built in MATLAB® and made use of the Image Processing Toolbox. It was run on a Dell XPS Studio PC with Intel® Core™ i7 CPU 360 3.2GHz processor and DDR3 8GB RAM with AMD Radeon™ HD6670 graphics card.

We tested our program with the highly diverse COREL1000 database [18]. It consists of 1000 images in which they are divided into 10 classes consisting of 100 images for each class. The classes are highly diverse, which consists of the classes: African People, Beaches, Buildings, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountains, and Food. From each class, 70 images are used for training (building the visual dictionary) and 30 images are used for query, totaling 700 images for training and 300 images for query.

The prototype consists of 2 main phases: Training Phase and Query Phase.

A. Training Phase

Training images from the first group are first fed into the SURF function. It will extract the interest points from each image with its respective 64 dimensions descriptors. The interest points will then clustered into k clusters using k-means algorithm, using Euclidean distance, with respect to their descriptors. For this experiment, we choose k = 100. We chose k = 100 because from our experiment, k = 100 have the best accuracy, precision, and computational time ratio. We could see the comparison of different k in term of accuracy and precision in Figure 2 below:

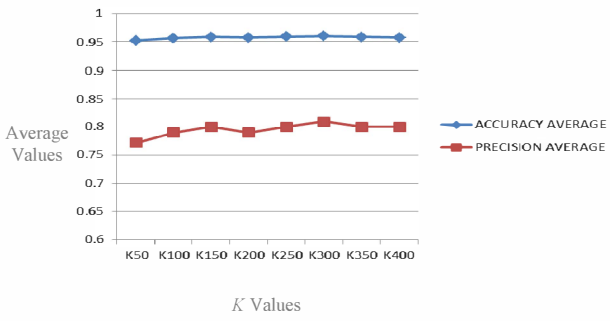


Figure 2. K Values Comparison Graph

Using our method, all values of k yields 0.96 and above in term of average accuracy, while in term of precision, all values of k yields above 0.75. But, the computational time increases significantly every time the value of k increases. For example, for k = 50 and k = 100, the training phase took approximately 6 hours of computational time while for k = 150, the training phase took 8 hours 30 minutes. For k = 300 which yields the highest precision, it took more than 12 hours for the training phase. Thus, we decided that it is not feasible to utilize k = 300 in our training and chose to use k = 100 instead. We could see from the result explained in Chapter IV that using k = 100, our method still outperforms the other methods.

We took the center of each cluster, chose it as the 'representative' of the cluster, and called it a visual word. Thus, we have a visual dictionary for the first group which consisted of 100 visual words.

This process was then repeated to the training images from the other groups. So in the end, we will have 10 visual dictionary, consisting of 100 visual words each, for the 10 groups from COREL1000 database.

We took the extracted features (interest points and descriptors) from the images in the training phase and calculated the Euclidean distance of each interest point with each visual words in its respective group visual dictionary and then clustered them according to the smallest distance (nearest neighbor). In other words, for each image; we mapped the features back to the group visual dictionary. For each cluster, we count the number of interest points clustered in it and produce a histogram that showed how many interest points are clustered for each visual word. This histogram is what we call 'Bag-of-Visual-Word Histogram' and it represents each image according to its group visual dictionary.

Our method is different from Hierarchical K-means [19]. Hierarchical K-means (HKM) is one of the variant of K-means clustering algorithm, which aims to classify variables into similar groups without prior knowledge of assigned groups; while our method proposed a novel dictionary building algorithm for BoVW to achieve better classification and retrieval with prior known classes, not for clustering variables. HKM could be employed within our proposed algorithm to replace K-means as the clustering algorithm. However, our experiment showed that K-means performs better than HKM in our case, which can be seen in Figure 3. Thus, we decided to use K-means instead of HKM.

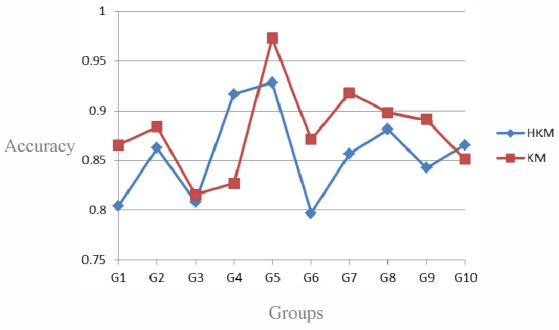


Figure 3. HKM and K-means Comparison Graph

B. *Query Phase*

When a user submitted a query image, interest points and descriptors will be extracted using the same SURF algorithm. It will then calculate the distance from each interest point in the query image to each visual word in the visual dictionary for the first group using Euclidean Distance. From each interest points the shortest distance is chosen and then summed up from all the interest points in the query image. This way, we have the minimum distance of the query image to the first group.

The process is then repeated to all other visual dictionary for the other groups. Once the process ends, the query image should have 10 minimum distances, representing the distance of the query image to each group. We will then choose the smallest minimum distance and classify the query image to the group with the smallest distance to the query image.

When a query has found its matching group, its features will also be mapped back to the codebook. The extracted interest points and descriptors will be then clustered to the visual words by calculating the distance using Euclidean distance and choosing the smallest distance of each interest point to each visual words. Then we produce a histogram again which showed how many interest points clustered for each visual word. This way the query image will have its own BoVW Histogram.

The query image BoVW Histogram will then be matched to the training images belonging in the same matching group to return the highest matches. The matching is done using the Histogram Intersection algorithm which was first introduced by Michael J. Swain and Dana H. Ballard [20]. Given two BoVW Histograms, their intersection is given by:

$$H_1 \cap H_2 = \sum_i \frac{min\{H_1(i), H_2(i)\}}{\max\{H_1(i), H_2(i)\}} \qquad (1)$$

The prototype will then return *N* highest number of matches (*N* number of matches could be determined by the user). Figure 4 shows the GUI prototype as the result of our experiment, it returned matching images for the Elephant query image (image 511 from COREL1000 database):

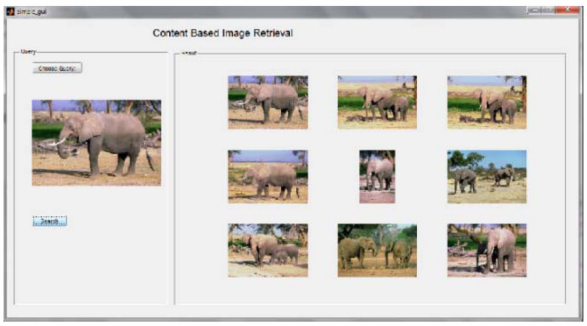


Figure 4. Prototype Result for the Query Elephant

IV. RESULT

This section presents all result of our experiment and comparison with other works. To denote the precision and accuracy of the retrieval and classification performance, we used the confusion matrix. Confusion matrix is a table used to evaluate the performance of machine learning classifier during supervised learning.

From the confusion matrix we could calculate the accuracy of the classification using the formula below [20]:

$$Accuracy = \frac{True\ Positives}{\frac{True\ Positives + False\ Negatives +}{True\ Negatives + False\ Positives}} \qquad (4)$$

True Positives is the number of images correctly classified to the correct group (e.g. query from Group 1 correctly classified by the system as Group 1) while False Negatives is the number of images from Group 1 that is incorrectly classified *not* as Group 1. True Negatives is the number of images that is not from Group 1 and correctly classified as other groups. And we could calculate the precision of the classification using the formula [21]:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \qquad (5)$$

False positives is the number of images that is not Group *n* but classified into Group *n* (e.g. query from Group 2 wrongly classified into Group 1). It is worth noting that we calculate accuracy and precision for each group, so each group will have its own True Positives and False Positives.

We compared our method, which is GBoVW (G-BoVW) with 3 other methods: Normal BoVW (N-BoVW) with global dictionary, Fuzzy Indexing BoVW (FI-BoVW) by Wassim Bouachir, Mustapha Kardouchi, and Nabil Belacel [22], and Weighted Histograms as input Mean-Shift and Gaussian Mixtures (WHMSGM) by Mohamed Ali Bouker and Eric Hervet [23]. FI-BoVW used SIFT as their chosen feature extractor and proposed a new weighting scheme using fuzzy representation to index image features for a more robust signature. WHMSGM modelizes the colors of an image as a set of 2D Gaussian distributions based on weighted color histograms. Beside color histogram, WHMSGM also made use of the color packets features.

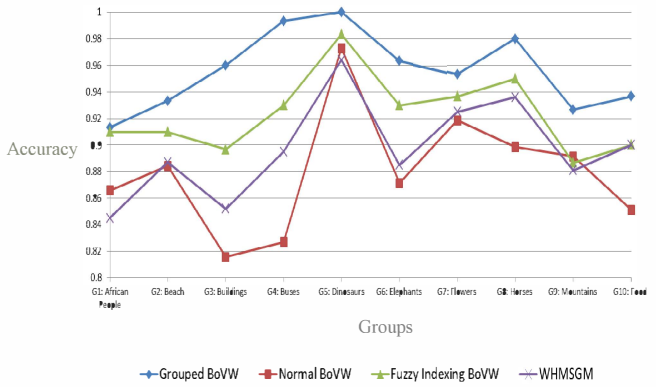


Figure 5. Accuracy Comparison Graph

From Figure 5, we could see that our method outperforms the other methods. We found that Group 5, which is Dinosaurs, achieved the highest accuracy which is 100% accuracy. The second highest accuracy is Group 4, which is Buses, with 99.33% accuracy. The lowest accuracy is 91.33% for the Group 1, which is African People. Our method's accuracy average comes down to 95.6% and it's higher than the Normal BoVW accuracy average (87.9%), Fuzzy Indexing BoVW (92.33%), WHMSGM (89.7%).

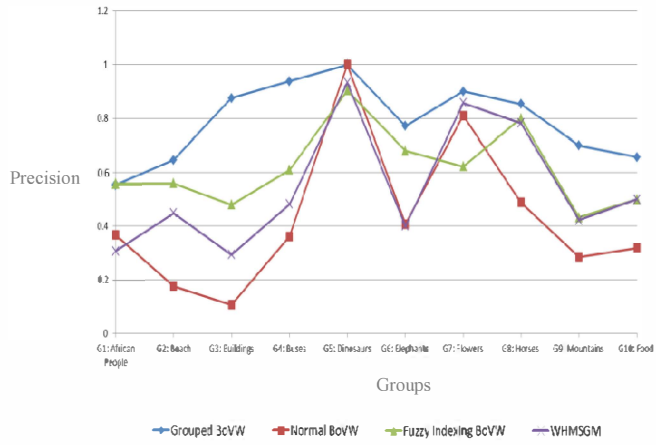


Figure 6. Precision Comparison Graph

Figure 6 above shows precision comparison. Same with the accuracy, Group 5 (dinosaurs) reached 100% precision. The lowest precision is Group 5 (African People) with 55.56% precision. Our method's precision average comes down to 78.97% and it's higher than the Normal BoVW precision average (43.23%), Fuzzy Indexing BoVW (61.49%), WHMSGM (54.25%). The chart comparisons show that our method outperforms the other methods in term of accuracy and precision.

We observed that the chosen feature extractor plays a big role in determining the precision and accuracy of the method. Our chosen method to extract image features, SURF, has been proven to be superior to other feature extractor used in FI-BoVW (SIFT) and WHMSGM (weighted color histogram and color packets). SURF is basically a shape-based feature extractor, thus images with clear objects excels in precision and accuracy for our method. This could be seen in the case of Group 5 (dinosaurs), Group 7 (flowers), and Group 9 (mountains) where the result of NBoVW and ours is superior. The aforementioned groups all have clear objects outlined in the image, thus making the precision and accuracy superior to other methods.

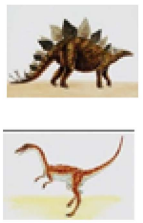
G5	G7	G9
		
G3	**G6**	**G10**
		
		

Table 1. Example of Images from each Group

Because of the chosen feature extractor, one weakness of it is if the image is cluttered with objects or contains more than one object, such as Group 3 (buildings), Group 6 (elephants) or Group 10 (food), the precision and accuracy will drop, as shown in the result of NBoVW. To accommodate this, we proposed the GBoVW. As can be seen from Figure 5 and Figure 6, our method with visual dictionaries for each group has proven to be more discriminative and presents higher precision and accuracy compared to other methods.

Thus, we can conclude that our method is biased toward shape, or images with clearly outlined object while the GBoVW made up for it by being more discriminative. Individual dictionary for each group presents a more accurate standardization for image comparison compared to global dictionary. We believe that for groups heavy in color, such as Group 7 (flowers) our result could be further elevated by incorporating color features in combination with SURF.

V. Conclusion And Future Work

In this paper we presented a new approach of building visual dictionary for the Bag-of-Visual-Words (BoVW) method. Our method created visual dictionary for each group in the COREL1000 database, as opposed to the global dictionary which normal BoVW employ. Compared to the the normal BoVW and a few other methods related to BoVW, our method

outperforms them in terms of accuracy and precision. Our GBoVW method is more discriminatory due to the individual group visual dictionary.

Our major challenge to the work is that our method is highly supervised. Highly supervised method means we need to determine the number of group before we perform classification.

For our future work, we would like to combine SURF features with some other methods, such as color histogram or color correlogram, which might produce even higher accuracy and precision.

REFERENCES

[1] John Eakins and Margaret Graham, "Content-Based Image Retrieval," JISC Technology Applications, University of Northumbria at Newcastle, October 1999.

[2] Michael J. Swain and Dana H. Ballard, " Color Indexing," International Journal of Computer Vision, 7:1, 11-32, Kluwer Academic Publishers, 1991.

[3] Shi-Kuo Chang and Sho-Hung Liu, " Picture Indexing and Abstraction Techniques for Pictorial Databases," IEEE Transaction of Pattern Analysis and Machine Intelligence, 1984.

[4] A. W. M. Smeulders, M. Worring, A. Gupta, R. Jain, "Content-Based Image Retrieval at the End of the Early Years," IEEE Transaction of Pattern Analysis and Machine Intelligence, 2000.

[5] X. S. Zhou and T. S. Huang, "CBIR: From Low-Level Features to High Level Semantics," Proceedings of the SPIE Image and Video Communicaitons and Processing, Vol. 3974, January 2000.

[6] Ying Liu, Dengsheng Zhang, Guojun Lu, Wei-Ying Ma, "A Survey of Content-Based Image Retrieval with High Level Semantics," The Journal of the Pattern Recognition Society, ELSEVIER, 2006.

[7] Herbert Bay, Tinne Tuytelaars, Luc Van Gool, "Speeded-Up Robust Features," Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346-359, EECV, 2008.

[8] David G. Lowe "Object Recognition from Local Scale-Invariant Features," The Proceedings of the Seventh IEEE International Conference on Computer Vision, Vol. 2, pp. 1150-1157, 1999.

[9] Christopher Evans "Notes on the OpenSURF Library," University of Bristol, 2009.

[10] Shihua He, Chao Zhang, Pengwei Hao, "Comparative Study of Features for Fingerprint Indexing," 16th IEEE International Conference on Image Processing (ICIP), pp. 2749;2752, November 2009.

[11] Maya Dawood, Cindy Cappelle, Maan E. El Najjar, Mohamad Khalil, Denis Pmorski, "Harris, SIFT, and SURF Features Comparison for Vehicle Localization based on Virtual 3D Model and Camera," 3rd International Conference on Image Processing Theory, Tools, and Applications (IPTA), pp. 307;312, October 2012.

[12] Luo Juan, Oubong Gwun, "SURF Applied in Panorama Image Stitching," 2nd International Conference on Image Processing Theory, Tools, and Applications (IPTA), pp. 495;499, July 2010.

[13] Anderson Rocha, Siome Goldenstein, Tiago Carvalho, Jacques Wainer, "Points of Interest and Visual Dictionary for Retina Pathology Detection," Instituto De Computacao, Universidade Estadual De Campinas, March 2011.

[14] Ke Gao, Shouxun Lin, Yongdong Zhang, Sheng Tang, Huamin Ren, "Attention Model Based SIFT Keypoints Filtration for Image Retrieval," Seventh IEEE/ACIS International Conference on Computer and Information Science, May 2008.

[15] P. Quelhas, F. Monay, J. –M. Odobez, D. Gatica-Perez, T. Tuytelaars, "A Thousand Words in a Scene," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No. 9, pp. 1575-1589, Sept. 2007.

[16] Tom Botterill, Steven Mills, Richard Green, "Speeded-Up Bag-of-Words Algorithm for Robot Localisation Through Scene Recognition," IEEE 23rd International Conference on Image and Vision Computing New Zealand (IVCNZ), November 2008.

[17] Anna Bosch, Xavier Munoz, Robert Marti, "A Review: Which is the Best Way to Organize/Classify Images by Content?," Image and Vision Computing, ELSEVIER, 2006.

[18] James Z. Wang, Jia Li, Gio Wiederhold, "SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 9, pp. 947-963, 2001.

[19] Kohei Arai, and Ali Ridho Barakbah, "Hierarchical K-means: an algorithm for centroids initialization for K-means." Reports of the Faculty of Science and Engineering, Vol. 36, No. 1, pp 25-31, 2007.

[20] Michael J. Swain and Dana H. Ballard, "Indexing via Color Histogram," IEEE Proceedings, Third International Conference on Computer Vision, pp. 390-393, December 1990.

[21] David L. Olson and Dursun Delen, "Advanced Data Mining Techniques," pp. 138, Springer, 2008.

[22] Wassim Bouachir, Mustapha Kardouchi, Nabil Belacel, "Improving Bag of Visual Words Image Retrieval: A Fuzzy Weighting Scheme for Efficient Indexation," IEEE Fifth International Conference on Signal Image Technology and Internet Based System, pp. 215-220, 2009.

[23] Mohamed Ali Bouker and Eric Hervet, "Retrieval of Images Using Mean-Shift and Gaussian Mixtures Based on Weighted Color Histograms," Seventh International Conference on Signal-Image Technology and Internet-Based Systems (SITIS), vol., no., pp.218,222, Nov. 28 2011-Dec. 1 2011.